Computers & Graphics 60 (2016) 93–106

Contents lists available at ScienceDirect

Computers & Graphics

journal homepage: www.elsevier.com/locate/cag

Special Section on SIBGRAPI 2016

Depth functions as a quality measure and for steering multidimensional projections

Douglas Cedrim [a], Viktor Vad [b], Afonso Paiva [a], M. Eduard Gröller [b,c], Luis Gustavo Nonato [a], Antonio Castelo [a]

[a] ICMC–USP, São Carlos, Brazil
[b] TU Wien, Austria
[c] VRVIS Research Center, Austria

ARTICLE INFO

Article history:
Received 29 February 2016
Received in revised form
23 August 2016
Accepted 23 August 2016
Available online 1 September 2016

Keywords:
Visual analytics
Depth functions
Non-parametric statistics
Dimensionality reduction
Quality measures

ABSTRACT

The analysis of multidimensional data has been a topic of continuous research for many years. This type of data can be found in several different areas of science. A common task while analyzing such data is to investigate patterns by interacting with spatializations of the data in a visual domain. Understanding the relation between the underlying dataset characteristics and the technique used to provide its visual representation is of fundamental importance since it can provide a better intuition on what to expect from the spatialization. In this paper, we propose the usage of concepts from non-parametric statistics, namely depth functions, as a quality measure for spatializations. We evaluate the action of multidimensional projection techniques on such estimates. We apply both qualitative and quantitative analyses on four different multidimensional techniques selected according to the properties they aim to preserve. We evaluate them with datasets of different characteristics: synthetic, real world, high dimensional; and contaminated with outliers. As a straightforward application, we propose to use depth information to guide multidimensional projection techniques which rely on interaction through control point selection and positioning. Even for techniques which do not intend to preserve any centrality measure, interesting results can be achieved by separating regions possibly contaminated with outliers.

1. Introduction

The importance of data analysis has grown tremendously in the last years. It has become a challenging task for many different reasons. First of all nowadays *data sources are ubiquitous* and are available *for a broader audience*, one of the reasons is that high-definition sensors have become less expensive (e.g., high-definition cameras, 3D scanners). Secondly, the *scalability* since the volume of user generated data in a small scale of time (e.g., hours) can easily achieve the range of gigabytes (or terabytes). At last but not least, the *complexity* of the data itself is also an important aspect to be taken into account. Some examples of such data are collection of images [1] and textual data [2], computational simulations [3], gene data [4] and so on.

While dealing with multidimensional, possibly high dimensional, data there have been many efforts in the visualization on providing techniques and tools to allow for data analysis. Common approaches are visual exploration through linked views [3,5,6] and multidimensional projections [1,7]. In order to improve the effectiveness of the analysis, *quality metrics* can be defined to aid a particular visualization, which allows for quantifying how much a particular visual design conveys relevant patterns of the data (e.g., cluster information). The works by Bertini et al. [8] and more recently Sedlmair and Aupetit [9] provide a comprehensive and systematic analysis of such quality metrics.

On the other hand, a common approach for data analysis comes from machine learning, which automatically generates hypotheses from the input data [10]. These hypotheses have many different uses. Commonly they delimit meaningful regions of the input space (i.e., identifying cluster regions) by defining decision boundaries between data from different categories or classes. Although algorithms that perform such operations in an automatic fashion have proven quite useful, they rely on the definition of a proper similarity metric among all pairs of the input data.

The idea of *proper* can vary with the type of data being analyzed. It is rather complex to define feature extractors, which are discriminative enough for a broad class of datasets. This is mainly due for two reasons: the *diverse* sources of the data (e.g., collection of images vs car engine design [3]) and because of *ambiguities*.

Multidimensional scaling techniques play a particularly important role in the context of such general datasets. They decrease the complexity of the analysis by reducing the dimensionality of the data. This enables, for instance, projecting the input data onto a visual space aiming to preserve constraints

http://dx.doi.org/10.1016/j.cag.2016.08.008
0097-8493/

(e.g., pairwise distances) as much as possible. Scatter plots have been used to effectively convey absolute and relative distances between points projected into the visual space, and to allow for interactions within exploratory environments [11].

Reasoning on spatializations through scatter plots involves two different aspects: the definition of objective quality measures (e.g., distance preservation) as well as subjective ones (e.g., user-specific metrics). One of the main challenges in this process is the possible mismatch between the objective measures and the user-made quality judgments [12,13]. As the human perception is highly based on pattern finding, the design of visual metaphors should provide ways of handling both aspects without much effort. In the literature, there are some works aiming to perform a quantitative analysis of different patterns that can be found on these spatializations. For example they analyze separation factors of clusters [12] or more general graph-based measures [14]. Wilkinson et al. [14] evaluate the presence of outliers as one important property to characterize scatter plots. They propose to use a minimum spanning tree to quantify the appearance of outliers, defined for the whole scatter plot rather than performing a point-wise analysis. Moreover, *outlier detection* itself is an active interdisciplinary area of research [15].

On the other hand, scalar fields defined over the data are interesting, because they allow for various types of analysis. For instance, finding patterns in different dimensions of the data can be handled by a visual inspection of the scalar field coordinate by coordinate. Also the scalar field can be analyzed quantitatively by using topological tools, such as persistent homology [16].

Data depth is a particular interesting scalar field which comes from order statistics and non-parametric multivariate statistical analysis. In order statistics no – or as few as possible – assumptions from the underlying data distribution are made beforehand [17,18]. It is tightly related to multivariate median estimation, since the latter does not have a single generalization from the unidimensional situation. In the multivariate setting, different generalizations for the median are provided by data depth functions [17].

Data depth functions convey the notion of centrality concerning the data. They also relate to methods of *extreme value analysis* in the outlier detection literature, as outliers can be seen as the least central points in the data. Such points might contain useful information about the data such as an abnormal behavior during their acquisition or synthesis; data with different underlying distributions mixed together, and abnormal patterns introduced while processing the data (e.g., multidimensional projection) [19].

Once the data depth distribution is calculated before and after some processing on the data (e.g., multidimensional projection), a global analysis through statistical tests could be done. Alternatively, a visual analysis provides a qualitative way of understanding how the depth distribution is given on the input space and how it has changed for individual data points. This also allows for including user knowledge in the process, since the way users perceive centrality on scatter plots could be taken into account, although this is out of the scope of this work.

Other statistics measures such as mean and variance, although widely used, can lead to misleading interpretations of the data distribution since they are easily influenced by outliers and also by non-symmetric data distributions [17].

The usage of the median as a more robust location estimation is considered an interesting alternative. It has an asymptotic *breakdown point* of 0.5, which is a robustness estimator. Only if half of the data were modified the location estimation would become completely corrupted [20]. To put this into perspective, the mean has a breakdown point of 0, i.e., a single outlier can completely modify the estimation. Robust statistical estimates have been shown to be successfully applied on different research areas outside statistics, such as image and geometry processing [21,22].

1.1. Paper outline and contributions

The paper is structured as follows: in Section 2 we discuss the literature of quality measures for multidimensional projections and also some approaches using statistics for multivariate analysis. In Section 3 depth functions are described more formally and the choice for a particular one is motivated. In Section 4 its use as a quality measure for multidimensional projections is discussed with some quantitative and qualitative experiments. Moreover a comparison with another quality measure is performed. In Section 5, we describe how data depth can be applied for control point selection and we explore some strategies for steering multidimensional projections. At last, we point out limitations, in Section 6.

Taking what has been previously described into consideration, the main contributions of this paper are:

- Using order statistics as a quality measure for spatializations of multidimensional data.
- A qualitative analysis tailored for visually inspecting candidate regions of multidimensional outliers.
- A quantitative analysis through a quality metric defined by individual point – data depth – variations.
- A framework for steering multidimensional projection techniques by different sampling strategies using data depth information.

2. Related work

One of the main challenges in information visualization is to increase the suitability of multidimensional data representation for data analysts [23]. Within this context, several quality measures have been proposed, in order to evaluate patterns in multidimensional data visualization. A common evaluation characteristic is the ability of the quality measure to identify clusters and relate that to how humans perceive scatter plots. For a comprehensive review of such quality measures, we refer to Albuquerque et al. [24] and to Tatu's thesis [13]. However, the proposed pipeline by Tatu [13] is rather different from ours, since it uses quality measures, selected by the user, to steer the multidimensional projection. Afterwards, appropriate dimensions are selected where the data is then projected onto.

Different lines of investigation evaluate spatializations by how much specific quantities are preserved after the projection procedure [25]. Etemadpour et al. [26] introduce the concept of *density-based motion* in order to evaluate the point density of clusters that can be lost when a multidimensional dataset is projected.

Three common categories of methods for measuring distortions are distance-based, topology-based (neighborhood), and perception-based methods.

Some distance-based approaches are not scale-invariant (i.e., standard stress measures), meaning that even identical spatializations might indicate totally different preservation situations. Moreover, even after a normalization procedure, datasets with outliers introduce a bias on the analysis. By definition, the distance of outlying points to non-outlying points is high. This affects how variations among non-outlying points are perceived, since they contribute less to the distance deviation measure [7].

Topological approaches mainly investigate the mismatch between neighborhoods on the input space and neighborhoods on the visual space [27,28]. This measure might be too strict, since small perturbations on the neighborhood of a point might considerably affect its neighborhood topology, although points still remain close. Some works try to address this problem.

In order to build a quality criteria for the assessment of dimensionality reduction, Lee and Verleysen [29] created a unified framework to represent several ranked-based distances using *K-ary* neighborhoods computed in high and low dimensional spaces. Venna et al. [30] provided a quality measure for an information-retrieval task called *NeRV* (neighbor retrieval visualizer), the measure relies on minimizing the cost of a query between the amount of missed instances and those wrongly retrieved. Heulot et al. [31] introduced *ProxiLens*, an interactive framework to detect and filter out false neighbors when visualizing projected data by using an adjustable radial neighborhood selection. The method proposed by Martins et al. [32] tries to alleviate this discontinuous behavior of neighborhood comparison by proposing a multi-scale approach. Moreover, other robust approaches can be used [16]. Another aspect of topological approaches is that they rely on the definition of a neighborhood graph, which might affect considerably the obtained results [33].

Although Aupetit [25] investigates how different multidimensional projection techniques distort the data, his approach focuses on distortion measures based on geometric entities (e.g., points, Delaunay edges). He also proposes a heuristic for analyzing similarities with respect to a reference point, which is also used by ProxiLens. On the other hand, in our work we explore a specific scalar field defined over the data, which is widely used in non-parametric statistics and outlier detection, namely the *data depth*. We also investigate the semantics associated with the distortion of data depth by a multidimensional projection technique. In the context of this work, possible spatializations are not limited to projecting the data onto linear combinations of the original data dimensions (e.g., as in PCA), but also allows for including non-linear multidimensional projection techniques.

Perception-based approaches investigate how the choice of a multidimensional projection affects the user perception and performance when executing typical tasks on visual layouts of the projected data. Albuquerque et al. [24] proposed a perceptual quality measure for scatter plots using machine learning. Firstly, the similarity between scatter plots is identified by the users and employed to train a multidimensional scaling embedding. Then, the scatter plots are ranked by the users according to their assignment to a given task. Etemadpour et al. [34] provided an approach to organize multidimensional data projection layouts driven by a user-centric task categorization. Recently, Etemadpour et al. [35] performed a comparative study to evaluate the user perception of multidimensional data projection layouts considering specific tasks related with distance preservation and outlier detection.

Some works have proposed to analyze multidimensional data using depth functions [17,36]. To the best of our knowledge, none of them aim specifically at analyzing how the choice of a particular multidimensional projection technique can affect the reasoning on the spatializations of the original data. In this direction, the work of Rousseeuw et al. [37] generalized the idea of a box plot to a two-dimensional setting by estimating a depth function based on convex-hull peeling operations. Initially depth functions have been proposed for analyzing points, and several approaches have been used in the statistics literature [17,38]. These statistical measures have proven quite useful for uncertainty visualization of ensembles [39–41]. Additionally, Potter et al. [42] propose a visual metaphor for summaries of different statistical moments, however their approach does not allow for a connection with spatializations.

3. Depth functions

In the one dimensional case, order statistics are quite important and have been used for a long time as they allow for computing important statistical measures (e.g., median, outliers). Its generalization to a multivariate setting is not as straightforward as for the mean. A component-wise median is a rather poor generalization, for instance [43]. Many different approaches for the multivariate median have been proposed in the statistics literature [44].

Depth functions play an important role in such a context, since the deepest point of a dataset can be taken as multivariate median. They can be viewed as a multivariate generalization of the one dimensional order statistics. The idea is to define a center-outward ordering of the data, allowing for extracting meaningful statistics on them, e.g., the multivariate median, as the most central point. In fact, depth functions provide one way to relate different statistical methodologies (e.g., order statistics, quantile methods) using a single non-parametric estimation [18].

Although there is a simple and clear ordering intuition behind it, from the most central to the least central, the notion of depth may vary depending on the proposed depth function. Later on in this section we show how this notion changes on some examples of symmetric and asymmetric multivariate data distributions.

More formally, let F_n be an empirical distribution of $X=\{\mathbf{x}_1,\ldots,\mathbf{x}_n\}$, sampled from a probability distribution F in $\mathbb{R}^m$, $m\geq 1$. The data depth is a way of measuring how deep a given point $\mathbf{x}\in\mathbb{R}^m$ is w.r.t. the underlying empirical distribution of the set X.

In general, depth functions should satisfy the following properties: affine invariance; maximality at the center; monotonicity relative to the deepest point; vanishing at infinity [45]. We are going to describe four different depth functions, which theoretically are able to handle multidimensional datasets.

One of the earliest data depth estimates is the *Mahalanobis depth function* (MHD). Its general idea follows the Mahalanobis distance, which relies on an anisotropic elliptical distribution of the data, approximated by the covariance of the data samples. It is defined as

$$MHD(F_n,\mathbf{x})=\left[1+(\mathbf{x}-\mu_{F_n})^T\Sigma_{F_n}^{-1}(\mathbf{x}-\mu_{F_n})\right]^{-1} \tag{1}$$

with μ_{F_n} and Σ_{F_n} being the mean and empirical (sampled) covariance matrix of F_n. Because of its simplicity concerning calculation and its simple intuition can be thought as an initial approach. However as it is based on non-robust estimates (e.g., mean and covariance), it can have a low breakdown point, imposing a serious limitation while dealing with outlying data, as one might see in Fig. 1.

An interesting aspect of the Mahalanobis depth estimation is the possibility of further extensions using kernel methods, which leads to the *kernel mapped Generalized Mahalanobis depth* (*kmGMHD*) [46]. The construction assumes that the data $\mathbf{x}_i\in\mathbb{R}^m$ is implicitly mapped into a Hilbert space $\mathcal{H}$ in which the data depth is then computed only by evaluating a kernel function on the input space data, a procedure known as kernel trick. This data depth function relies on the computation of nontrivial dot products in the Hilbert space $\mathcal{H}$. Using the kernel trick, it is possible to compute these dot products in $\mathcal{H}$ only in terms of evaluating a kernel function k on the input data. The intuition is this procedure allows for capturing possible non-linearities on the data by the action of a kernel function. The depth is defined as

$$kmGMHD(x)=\left[1+\sum_{i=1}^{r}\frac{((\hat{k}(\mathbf{x},\mathbf{x}_j))^{j=1,\ldots,n}\mathbf{u}_i)^2}{\lambda_i^4}\right]^{-1} \tag{2}$$

where $\hat{k}(\mathbf{x},\mathbf{x}_j)=k(\mathbf{x},\mathbf{x}_j)-\frac{1}{n}\sum_{l=1}^{n}k(\mathbf{x},\mathbf{x}_l)-\frac{1}{n}\sum_{l=1}^{n}k(\mathbf{x}_j,\mathbf{x}_l)+\frac{1}{n^2}\sum_{l=1}^{n}k(\mathbf{x}_j,\mathbf{x}_l)$ and $\{\lambda_i^2,\mathbf{u}_i\}$ are the r non-zero eigenpairs of the kernel matrix $k(\mathbf{x}_k,\mathbf{x}_l)$ evaluated in all n points.

Kernel functions can be defined for various entities, like points, texts, image patches, graphs. For a comprehensive review of kernel

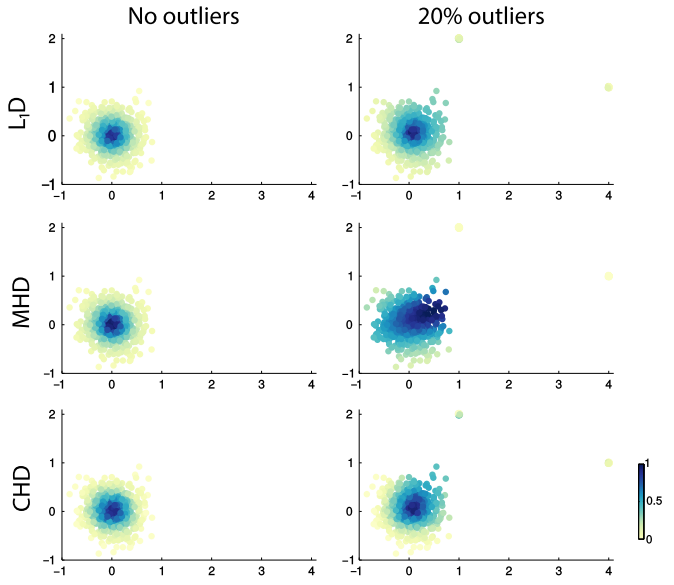

Fig. 1. Robustness evaluation in the presence of outliers, color coded by data depth computed in the two-dimensional input space. The closer to one, the more central a point is (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

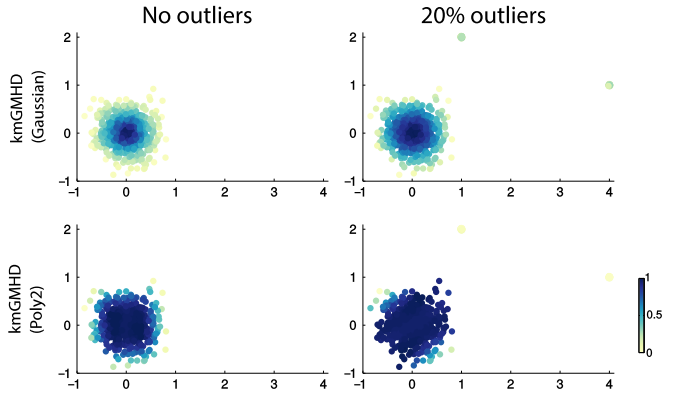

Fig. 2. Robustness evaluation on the presence of outliers in a two dimensional space, color coded by data depth using kernels. The closer to one the more central the point is (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

functions the reader is referred to Hofmann et al. [47]. In the literature, commonly used analytic kernel functions are the Gaussian kernel $k(\mathbf{x},\mathbf{y})=\exp(-\|\mathbf{x}-\mathbf{y}\|^2/2\sigma^2)$ and the polynomial kernel $k(\mathbf{x},\mathbf{y})=(\langle\mathbf{x},\mathbf{y}\rangle+1)^p$. There is a well defined machinery for constructing new kernel functions from existing ones. This allows for arbitrarily many kernel combinations and analyses. It is, however, not clear whether these are still going to define data depths.

We have repeated the same tests shown in Fig. 1 using kmGMHD. The possibility of changing the behavior of the depth by changing the kernel function introduces additional flexibility, but requires the understanding of the effect of the kernel choice. Fig. 2 shows that the kernel choice can significantly modify the depth distribution. For the polynomial kernel (Poly2) we have chosen degree $p=2$ and we noticed that changing this parameter did not affect the result considerably. With the Gaussian kernel, the results were obtained using Silverman's rule of thumb, $\sigma=1.06\hat{\sigma}n^{1/5}$, where $\hat{\sigma}$ is the standard deviation estimated from the samples [48]. However, a problem with this estimation is its sensitivity to outliers, which is not helpful for our goal. In Fig. 3, we show how good it captures nonlinearities in the data.

Hoffmann [49] reconstruction error is a kernel-based measure for novelty detection. It is tightly related to the distance of the data to its estimated PCA plane in Hilbert space $\mathcal{H}$ (kernel PCA). While using Gaussian kernel functions, this measure is tightly related to Parzen-window estimates, a widely known non-parametric density estimator.

Among several depth functions studied in the computational geometry literature, the *Convex hull depth* (*CHD*) is rather appealing because of its simplicity of being understood and computed. The basic idea is to start computing the convex hull of the input data. All points lying on the hull define the lowest depth of the data. All of these points are discarded from the computation and a new convex hull is computed, defining the next depth contour. This peeling process is repeated until the innermost layer is found. Although it is a good depth estimation, as one can see in Table 1, it can become infeasible in higher dimensions.

The multivariate L_1-median (geometric median, spatial median) is the theoretical solution of the Fermat–Weber location problem. Given the samples X (as before), and weights for the samples $W=\{w_1,\ldots,w_n\}$, the task is to find a point $\mathbf{y}$, which minimizes the weighted sum of the distances between $\mathbf{y}$ and the samples. Namely $\mathbf{y}$ is found as

$$\mathbf{y}=\underset{\mathbf{x}}{\operatorname{argmin}}\sum_{i=1}^{n} w_i\|\mathbf{x}-\mathbf{x}_i\|_2 \tag{3}$$

In the general case, a numerical solution is estimated by a simple iterative method, i.e., the Weiszfeld algorithm. Vardi and Zhang [50] proposed an improvement in the cases in which the estimated median is sufficiently close to one of the samples. Moreover, they define a depth function based on the proposed optimization procedure, named $L1$-Depth (L_1D).

The intuition of L_1D comes from the fact that $\mathbf{y}$ is the geometric median, so it can be considered as the most central point based on

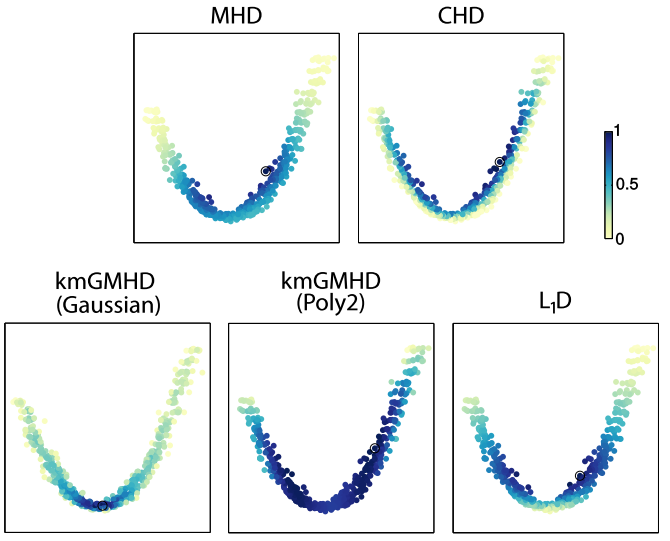

Fig. 3. Depth computation on a parabola-shaped dataset. The black circle illustrates the most central point according to the depth function used.

Table 1
Robustness and asymptotic complexity of different depth functions.

Data depth	Outlier robustness	Computational complexity
MHD	No	$O(n+m^3)$
kmGMHD	No	$O(n^3+m)$
CD	No	$O(n\log\ n+n^{\lfloor\frac{m+(1-m\bmod 2)}{2}\rfloor})$
L_1D	Yes	$O(nm)$

the samples, and it minimizes Eq. (3). Vardi and Zhang [50] have exploited the fact that the minimization function for Eq. (3) is continuous in $\mathbf{x}$. They use its direction of minimization, such that with given X and W the L_1D is defined as follows:

$$L_1D(\mathbf{x}) = 1 - \frac{\max(\mathbf{r}(\mathbf{x}) - \mathbf{w}(\mathbf{x}), 0)}{\sum_{i=1}^{n} w_i} \tag{4}$$

where

$$\mathbf{r}(\mathbf{x}) = \left\| \sum_{\mathbf{x}_i \neq \mathbf{x}} w_i \frac{\mathbf{x}_i - \mathbf{x}}{\|\mathbf{x}_i - \mathbf{x}\|} \right\|_2 \tag{5}$$

and

$$\mathbf{w}(\mathbf{x}) = \begin{cases} w_k & \text{if } \mathbf{x} = \mathbf{x}_k,\ k = 1 \ldots n \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

Although the weights introduce a flexibility in adjusting the importance of the points, in this paper we use $w_k = 1, k = 1 \ldots n$, i.e., giving all points the same weight.

In Fig. 1 we introduced outliers in order to assess how this affects data-depth distributions for non-robust depth estimates (e.g., MHD) compared to a robust one (i.e., L_1D). This shows that even after the addition of 20% of outliers, the depth distribution using L_1D is almost unchanged. Also, CHD does perform well under the presence of outliers.

In Fig. 3 five data depth function estimates are employed, illustrating their behavior in a two dimensional scenario, on a manifold-like point distribution.

The complexity of computing depth functions increases with the dimensionality of the input space. This imposes a serious limitation on the applicability of many depth functions for high-dimensional datasets. For instance, convex hull computations become computationally prohibitive already for datasets with 200 points in a ten-dimensional space, which is quite restrictive. Additionally, for most of the analyzed datasets, the covariance matrix estimated for the Mahalanobis depth has either shown to be not invertible or with an ill-conditioned behavior, and it seems to require a regularization procedure [51]. Kernel-based depth scales well with the dimension of the data, since its complexity only depends on the number of instances, making it suitable for datasets where the dimension $m \gg n$ (e.g., datasets of images). However, it involves the selection of a kernel function and parameter tuning procedure. Furthermore, for general multidimensional datasets it is not immediately apparent how this kernel choice affects the depth and it requires a more in-depth investigation. Some heuristics approaches already used for parameter tuning, like Silverman's [48], can be a good starting point for such an analysis.

In this work we focus our analysis on the L_1 data depth. Its computational complexity allows for practical computations and it locally preserves the data-depth distribution, even if the dataset is contaminated by a high number of outlier points (see Fig. 1).

Although it is not the primary focus of this work, other statistical quantities (e.g., kurtosis) can be derived from the data depth estimation. It has been shown that taking into account some of these data-distribution inherent characteristics while performing a multidimensional projection can significantly modify the result in the visual space and its understanding [17,52].

4. Depth as a quality measure

In order to describe the usage of data depth as a quality measure, we have set up some questions to motive our design choices for the analysis. More specifically these are:

Q1: *How to visualize data depth in multidimensional datasets*?

It is possible to find in the literature lots of works which make usage of depth functions to visualize entities in both 2D and 3D spaces (e.g., depictions of ensembles) [41] and also on graphs [53]. However, no visual metaphor aims specifically at visualizing data depth defined over points in a multidimensional space with a dimension higher than three.

We decided to address this problem by a simple approach of color coding the data depth field, namely D^m, computed on the input space $\mathbb{R}^m$ and directly depicted in the spatialization. Nevertheless we use a multi-hue, monotonically varying luminance color map from ColorBrewer [54], with the goal to continuously vary the perceived data depth values. The color mapping ranges from dark blue for the highest depth value (most central point), passing through light green for intermediate values, and going to light yellow for the least central value. The color map refers to the depth computed on the input space $\mathbb{R}^m$.

The experiments using such a strategy are described in Section 4.1 and the results can be seen in Figs. 4–7.

Q2: *How do Multidimensional Projection (MP) techniques modify depth after the spatialization*?

Once we have computed the data depth on the original space, we repeat the process for the data projected into the visual space $\mathbb{R}^2$, which gives the scalar field D^2. This defines two scalar fields over the data, both with ranges in the interval $[0, 1]$.

A simple approach to investigate how the centrality values changed during the multidimensional projection process is to calculate point-wise differences between both scalar fields, as there is a one-to-one relation between them. This produces a difference scalar field D^d with range on $[-1, 1]$.

The values in this range have an interesting associated semantic. If we define the difference scalar field

$$D^d = D^m - D^2, \tag{7}$$

the extreme values of D^d are going to convey the following behavior: a value closer to 1 (orange) indicates that a central point in the original space has been moved towards a peripheral region (i.e., low depth value) in the visual space. Such behavior defines a *False Peripheral Point* (FPP), since it is peripheral only in the visual space. Conversely, the closer the value is to -1 (purple), implies a peripheral point in the original space has been moved towards a central region (i.e., high depth value) in the visual space. We call such a point a *False Central Point* (FCP). Points that did not change their depth are going to have a neutral color, conveying their neutral behavior.

These definitions are closely related to the notion of False-Neighbors and Tears, by Lespinats and Aupetit [28] using the data depth instead of neighborhood information. False neighbors are points not in the same neighborhood in the input space, but neighbors after the multidimensional projection. Conversely, tears are points which are neighbors only in the input space.

The occurrence of FPPs and FCPs is of great importance, since it can indicate whether possible outlier points from the input space have been mixed together with non-outlier points after the multidimensional projection. To the best of our knowledge, none of the proposed techniques in the literature address this question directly. In order to assess how this strategy conveys such behavior, it has been applied for data with outliers, as described in Section 4.1, together with approaches to avoid clutter as proposed in Section 4.3. Some results are given in Fig. 9a, b and c.

Q3: *In which regions does one find good candidate points for steering a MP*?

Once the centrality values have been computed, they can be used for subsampling the dataset in the original space, according to its depth distribution. This process defines control points with an associated semantic (i.e., its centrality in the data), which allows for steering multidimensional projection techniques that

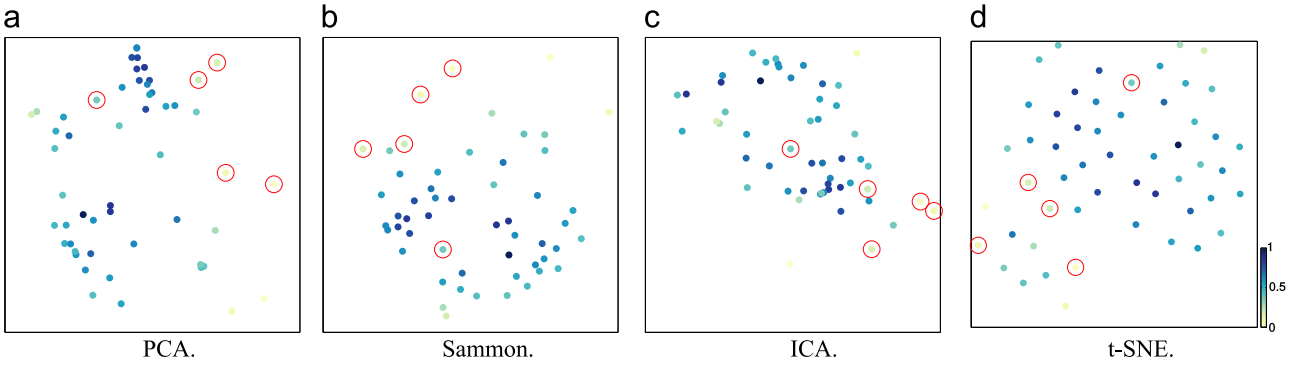

Fig. 4. L_1D depth computed on the Parkinson dataset with four multidimensional projection techniques. The data depth values are color coded with outliers highlighted in red (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

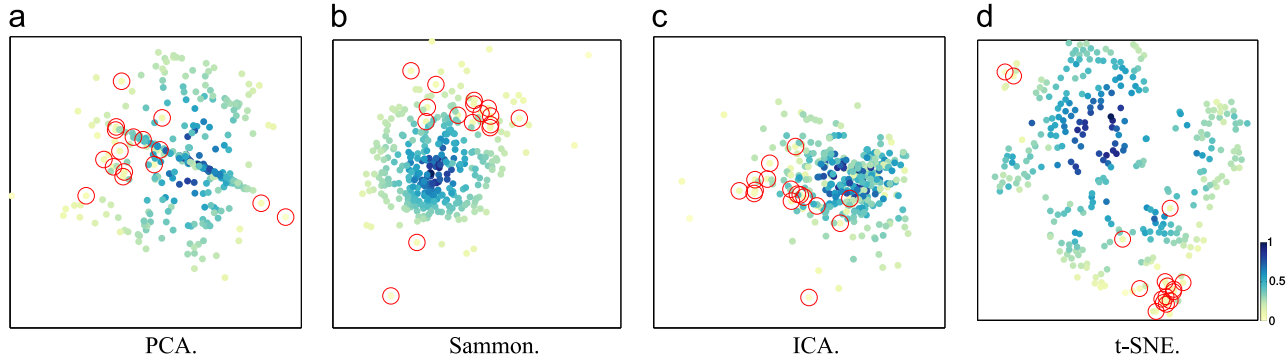

Fig. 5. L_1D depth computed on the Stamps dataset with four multidimensional projection techniques. The data depth values are color coded with outliers highlighted in red (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

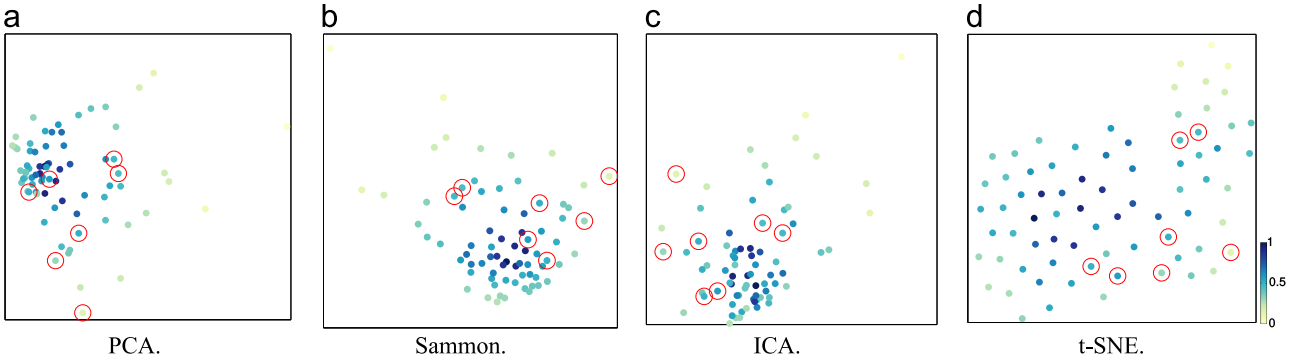

Fig. 6. L_1D depth computed on the Hepatitis dataset with four multidimensional projection techniques. The data depth values are color coded with outliers highlighted in red (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

rely on control point selections. In Section 5 we propose different strategies for this sampling process.

4.1. Experiments – qualitative evaluation

DD-plots have been used by statisticians to compare one data depth distribution against another one [17]. It is defined by a two-dimensional scatterplot where each point coordinate is its depth value in one of the distributions. This means that for two identical distributions, all points in a DD-plot lie on the line $y=x$, similar to the visual stress comparison shown in Joia et al. [1], and widely used in the information visualization literature. Although the analysis of depth changes becomes straightforward, as it consists of checking whether a point lies above or below the line $y=x$, it lacks an association with the spatializations of the original data. In contrast, our qualitative analysis relies on the strategies described together with the questions Q1 and Q2, where the depth changes are associated with color variations in the spatialization.

Since there is no multidimensional projection technique available, which aims to explicitly preserve centrality measures, we have decided to vary the choice of technique depending on the underlying assumptions on the data.

The first technique we have chosen is the Principal Component Analysis (PCA), since it has been most frequently used by practitioners [55]. It relies on finding the directions of maximal variance among the data, which is another statistical estimate. The second technique is Sammon mapping [56], which minimizes the inter-point distances in a non-linear fashion. The third one is the Independent Component Analysis (ICA) [57], which is targeted at non-Gaussian distributions of the data. The last one is t-SNE [52], which is based on probabilities defined on the points. It minimizes the difference between probability distributions in the original space and the visual space, and tends to preserve the local structure of the data.

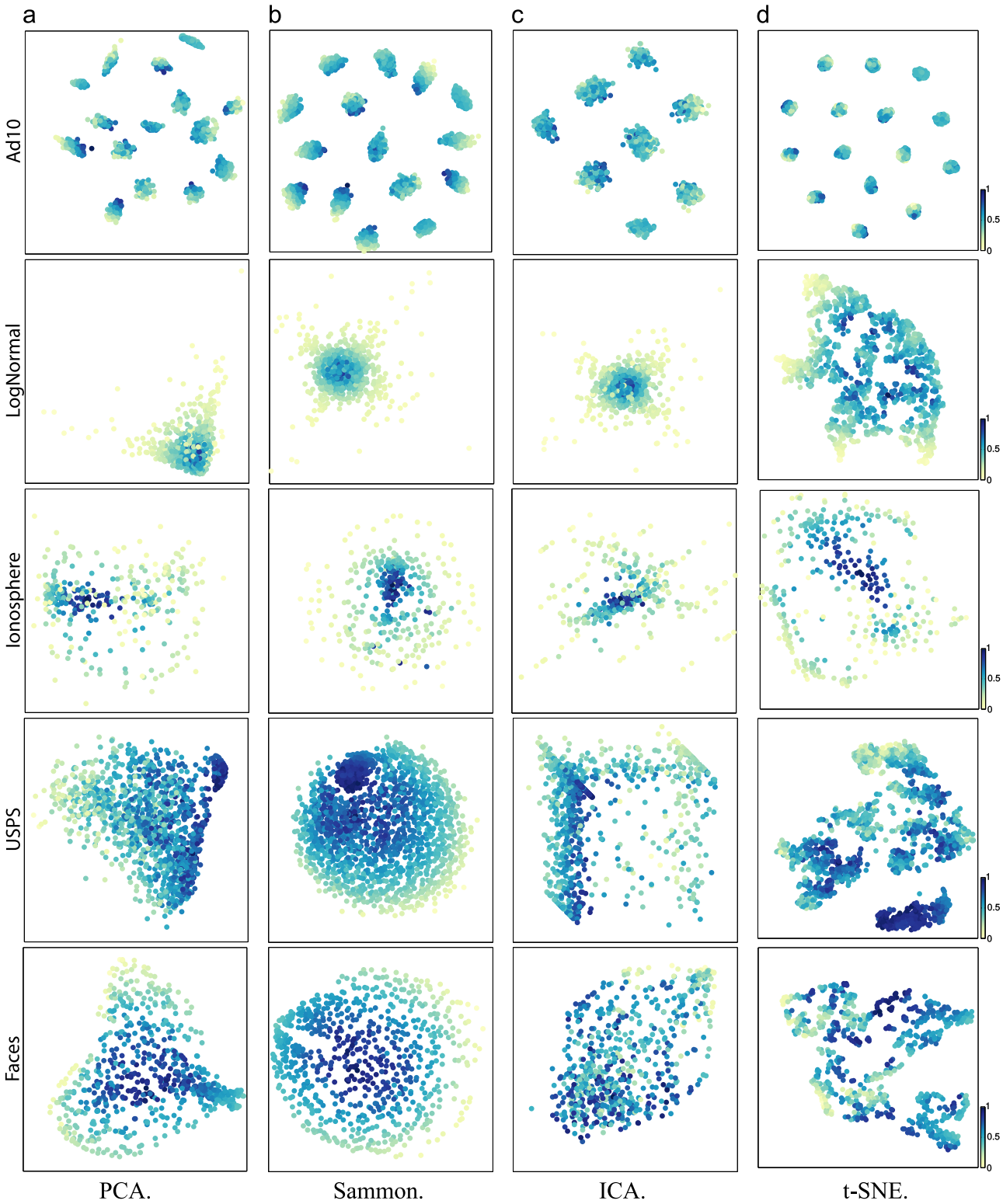

Fig. 7. L_1D computed on five different datasets with four multidimensional projection techniques (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

Note that any multidimensional projection technique which does not rely on control point selection could be used in this step.

4.1.1. Data with outliers

We have conducted experiments with three different datasets contaminated with outliers, which have been defined by domain experts for evaluating the performance of outlier detection techniques [15]. The datasets were chosen according to the increasing complexity of having their outliers detected. The datasets Parkinson, Stamps and Hepatitis were taken from the database provided by Campos et al. [15].

The *Parkinson dataset* is composed of medical data from 53 persons, each of the data instances with 22 dimensions.

People suffering from Parkinson's disease, i.e., 5 entries, are marked as outliers, which amounts to 10% of the dataset. Fig. 4 illustrates the behavior of the chosen multidimensional projection techniques on this dataset. In all techniques the outliers, highlighted in red, have a relatively low depth. Even in a rather simple dataset, regarding outlier detection, MP techniques still mix outliers with points with high depth values, after the projection.

The second dataset, *Stamps*, is composed of 325 color stamps classified as genuine stamps or forged ones (e.g., photocopied ones). The latter are considered to be outliers of the dataset and they make up around 5% of the dataset, which means 16 outliers. Data points are characterized by 9 different geometrical and color features, such as minimum bounding box, aspect ratio and pixel density. Fig. 5 shows that the outliers have low depth values, according to the L_1D. Similar to the previous experiment, they were mixed with points with high depth values after the multidimensional projection, with PCA and ICA producing the worst results.

The third dataset consists of 74 patients suffering from *Hepatitis*, and their corresponding predictions whether they are going to survive or going to die (i.e., outliers). This dataset has 19 attributes and 7 outliers (10%). From the three datasets, this is the most complex one to have its outliers detected. Fig. 6 also reflects these characteristics since outlier points do not necessarily have the lowest depth values.

We have performed other experiments aiming to check how the MP techniques perform on different kinds of datasets. The less central regions might contain outliers, although with no ground truth as used in this experiment.

4.1.2. Normally distributed data

The artificial dataset *AD10*, which is depicted in the first row of Fig. 7, is constructed as follows: fifteen clusters are generated following a Normal distribution $\mathcal{N}(0,1)$, and are placed at random vertices of a hypercube in a ten-dimensional space. The main purpose of this experiment is to assess how data depth and multidimensional projection techniques perform on several different clusters with a well known structure (i.e., being normally distributed).

Although t-SNE produces more compact clusters, it does not avoid mixing together points with quite different values of centrality, inside each cluster. The depth distribution in the Sammon mapping result preserves better than this distribution in comparison with the other techniques. This visual analysis is confirmed by the quantitative evaluation, as shown in Table 2.

4.1.3. Non-Gaussian data

Although a common assumption is to rely on Gaussianity of a data distribution, this might not be its underlying characteristics. Examples of skewed data distributions (e.g., log-normal ones) can be found in many different fields in science: geology, human medicine, microbiology, atmospheric sciences, social sciences, and economics [58]. In order to evaluate the behavior of Non-Gaussian distributed data we have created a synthetic *LogNormal* distributed dataset, following a $\ln \mathcal{N}(2000, 0.7)$ distribution, in a five-dimensional space.

As becomes apparent in the second row of Fig. 7, ICA performs quite well for this type of data, preserving the overall structure of the points. Is expected it to outperform other techniques, as is suited for such kind of datasets (i.e., non-Gaussian). However, it does not avoid mixing points with low and high depth values, illustrated in the region containing most central points. The results obtained using Sammon mapping technique still produces the best result regarding data depth preservation, while keeping the point distribution similarly.

Table 2
Data-depth distortion measured from D^d. The lower the value, the higher the depth preservation by the multidimensional projection technique (best results are shown in bold). Rows are grouped according to the assumptions on the dataset.

Dataset (DS)	n	m	PCA	Sammon	ICA	t-SNE
DS1: Parkinson	53	22	1.59	**1.51**	2.10	2.05
DS2: Stamps	325	9	3.09	**1.58**	3.95	2.84
DS3: Hepatitis	74	19	1.13	**0.85**	1.59	1.71
DS4: Ad10	1499	10	8.22	**7.39**	7.81	7.71
DS5: LogNormal	999	5	5.76	**4.22**	5.46	4.59
DS6: Ionosphere	349	34	3.62	**1.76**	3.34	3.17
DS7: USPS	1457	256	9.20	**6.89**	9.11	8.64
DS8: Faces	697	4096	4.94	**4.35**	5.83	5.48

4.1.4. Real-world data

The *Ionosphere* results from radar data of free electrons in the ionosphere. It is composed of two values per pulse of the processed electromagnetic signals, with 17 different pulses, i.e., giving a 34-dimensional dataset. The goal is to seek for evidence of some structure in the signals, classified as good or bad accordingly [59].

As can be seen in Fig. 7, the preservation of depth is nearly the same for all techniques, with Sammon and t-SNE technique producing the two best results (see Table 2).

4.1.5. High-dimensional data

The first high-dimensional data experiment, illustrated in Fig. 7, is the US Postal Service (*USPS*) digits dataset [59]. It is composed of images of digits with 10 different classes, from 0 to 9.

The second experiment, i.e., with the *Faces* dataset, illustrated in Fig. 7 is defined by 697 instances of 64×64 grayscale images of faces with different poses and lighting conditions [60]. Each image is represented by its vector form defining points in a 4096-dimensional space.

Although in this experiment PCA does preserve the notion of centrality, Sammon mapping produces the best results as compared to the other techniques. t-SNE still retains the continuous variation of the depth, defined on the input data. This is clearly not the case for ICA, which mixes together central points with peripheral points of the input space.

4.2. Experiments – quantitative evaluation

The difference scalar field D^d, defined for visually inspecting the behavior of the data depth, straightforwardly specifies a measure of data depth preservation. We construct a vector $\mathbf{s}$ of n entries defined by the D^d difference scalar field evaluated at each point $\mathbf{x}_i$ from the dataset, as follows

$$\mathbf{s} = \left(D^d(\mathbf{x}_1),\ D^d(\mathbf{x}_2),\ \ldots,\ D^d(\mathbf{x}_n) \right) \tag{8}$$

and we define the *data-depth distortion* as the L_2 norm of the vector $\mathbf{s}$, such that the closer to zero the lower the distortion on the depth. Also an upper bound for the depth distortion is $\sqrt{n}$, since the maximum absolute change in depth for a single point is not greater than one. The upper bound would be reached in the hypothetical situation where there would be a maximal depth change at each of the n points of the dataset. Table 2 quantifies the data-depth distortion for the various experiments. An interesting result is the overall very good performance of Sammon mapping in all the experiments.

We have implemented the data depth L_1D using MATLAB®. All experiments have been performed on an Intel Core i5-5200 CPU with 8 GB of RAM. Table 3 shows the computational timings for D^m using L_1D.

4.3. Clutter avoidance

In order to avoid cluttering in visualizations, as given in Fig. 7, one could use a strategy similar to Aupetit [25]. Their approach is based on partitioning the 2D space by a Voronoi diagram of the projected point set. Fig. 8 shows that regions with a higher density of points produce more Voronoi cells with smaller areas. Nevertheless, the approach is still well suited for analyzing the data-depth preservation. High density regions with color discontinuities, as illustrated in Fig. 8a, indicate an undesired behavior, in which points with quite different depth values were mixed together. Conversely, the continuity, as illustrated in Fig. 8b, shows that the projection has been able to locally preserve the data-depth distribution.

A possible drawback of such an approach is that the size of the cells might convey a misleading notion of importance, as compared to color discontinuities. Additionally, as shown in Fig. 8c and d, the impact of FPP and FCP becomes region-based instead of point-based, which could benefit projections with only few points.

4.4. CheckViz comparison

In general, as shown in Fig. 9, the results obtained using our approach are quite promising to convey data-depth preservation. There are interesting connections with the CheckViz approach [28], such as that regions of high distortion in our approach are also regions of high distortion in the CheckViz approach.

During the experiments the CheckViz scaling parameter σ has been selected as the average distance between each point and its fifth neighbor in the input space, a rule of thumb suggested by the authors.

The behavior of our approach is that outlying points, highlighted with red circles, are either kept in peripheral regions after the projection, which will produce gray Voronoi cells, or they are moved towards central regions, which will define purple FCP Voronoi cells. This makes both types of cells candidates for containing outliers, which can be confirmed in Fig. 9a and b. However, the same does not hold for some cells in Fig. 9c, which implies that the estimated depth value in the original space was moderate, not corresponding to what is expected from an extreme value outlier data. This is not a limitation of the proposed technique but of the chosen depth function to properly identify outliers as points with low depth values.

On the other hand, the expected behavior of the CheckViz approach is that the outlier points are either kept in peripheral regions, defining white Voronoi cells, or they are mixed together with more central points, defining false neighbors and producing purple Voronoi cells. In the same sense as our technique, both types of cells are candidates to contain outliers, and this behavior can be seen in Fig. 9e and f. However, Fig. 9d shows that outliers can also be classified as tears, making more difficult to analyze their behavior by using CheckViz approach.

Another interesting aspect can be seen in Fig. 9b and e, exposing an important distinction on the kind of performed analysis. More specifically, our analysis focuses on two regions $A: [-0.01, 0.05] \times [0.03, 0.10]$ and $B: [0, 0.07] \times [0.10, 0.16]$. CheckViz visually indicates almost the same distortion by the mapping in both regions A and B. Our approach reveals they have quite different behaviors: (a) region A contains peripheral points from the input space which are still peripheral after the projection, visually encoded with a neutral cell color; (b) region B contains central points from the input space which are now peripheral after the projection, according to the data depth estimation, which produces cells with orange colors.

These two different behaviors may have a direct impact on understanding the effect of the projection on the data. For instance, using our approach, one could spot regions with possible outliers by identifying less central cells with neutral colors, while the same cannot be provided by the CheckViz approach. User studies would be interesting to investigate the impact of such an encoding on the analyst's perception of the projection.

Table 3
Computational timings in seconds for computing D^m using L_1D.

Dataset	DS1	DS2	DS3	DS4	DS5	DS6	DS7	DS8
Time (s)	0.009	0.14	0.013	2.78	1.23	0.17	6.41	11.06

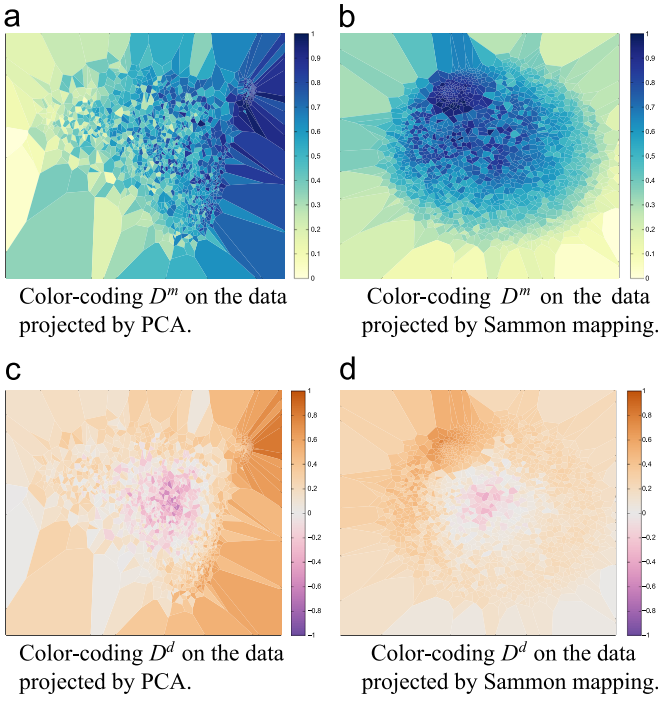

Fig. 8. Clutter reduction by constructing Voronoi diagrams of spatializations of the USPS dataset.

5. Steering multidimensional projections

Some multidimensional projection techniques rely on the selection of some points in the original space (i.e., control points) and to position them in the visual space, guiding the projection process to a certain extent. It allows the user to steer the projection by selecting points with an associated semantic (e.g., centroid of a cluster, outliers) and by carefully positioning them in the visual space. The goal is that the overall projection follows the control points as much as possible. The process of selecting appropriate control points and properly positioning them is highly dependent on the desired task. For instance, one could use class information to position points of different classes in distinct regions, aiming to improve the inter-class separation of the data.

Since computing depth functions defines a scalar field on the input data, in which the sorted values describe a notion of center-outwards ordering, one can use this information to steer the multidimensional projection. The centrality information can be

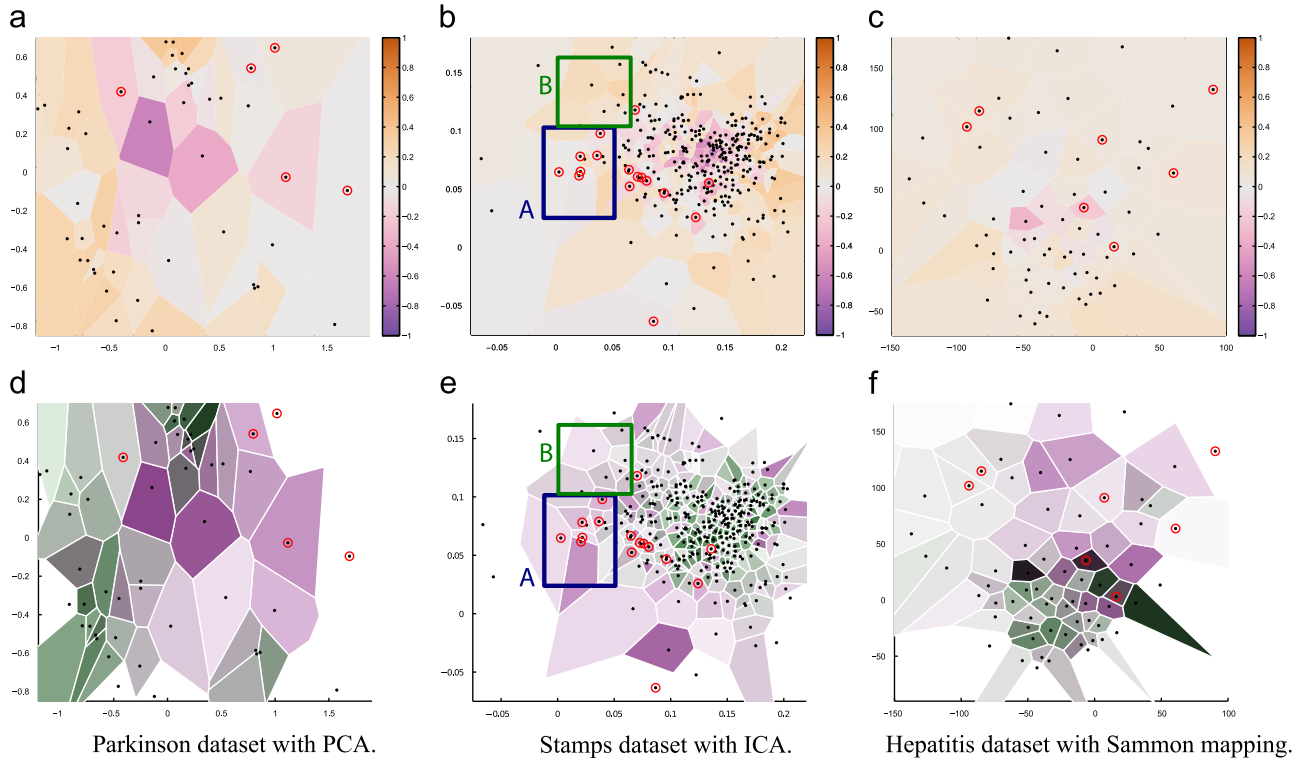

Fig. 9. On top row, the FPP (orange) and FCP (purple) regions, defined in Section 4. At bottom row, CheckViz tears (green) and false neighbors (purple) (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

used, so that after the projection outliers lie in distinct regions than points with large depth values.

In order to explore the centrality information, we have implemented two strategies for sampling the input space and for positioning control points in the visual space. We compared them to a randomly control point selection and placement. Both strategies aim to preserve data depth by sampling the visual space in a radial fashion, because of its known data depth distribution.

Many multidimensional projection techniques rely on minimizing distance distortions between input and visual spaces. Therefore, we calculate how incorporating depth information in the control point selection affects such measure, defined as the *stress*:

$$stress = \frac{\sum_{ij}(d_{ij} - \overline{d}_{ij})^2}{\sum_{ij} d_{ij}^2} \tag{9}$$

where d_{ij} and $\overline{d}_{ij}$ are the distances between points $\mathbf{x}_i$ and $\mathbf{x}_j$ in the original space and in the visual space, respectively.

The performance of each strategy is shown in Table 4, comparing its stress measure and the data-depth distortion, defined in Section 4.2, evaluated on five random experiments each. Moreover, for projecting in the visual space, we have chosen the Local Affine Multidimensional Projection (LAMP) technique [1] because of its effectiveness on preserving distances locally.

In order to minimize the effect of scale differences on the stress, the control points are mapped to the visual space inside a 2D-disk of diameter d_{max}, where d_{max} corresponds to the maximum distance between two points on the original space.

5.1. Random sampling (RS)

A straightforward sampling strategy is to randomly select points in the original space and place them at a random position inside the 2D-disk of diameter d_{max}. The idea behind this strategy is to investigate how the LAMP technique performs if the control points are chosen at random, without taking into account any further information about the data (e.g., data depth). In Fig. 10 projections obtained for the Stamps dataset are shown, varying the number of control points.

Although it is a possible strategy for selecting control points to project the data into the visual space, the randomness of control point selection and positioning lacks a more informative semantic, specially regarding the depth. This strategy just spreads out randomly the selected control points on the plane. However, the distance deviations of each projected point with respect to the control points is locally minimized, which is a characteristic of the LAMP technique.

Table 4
Quantitative analysis of the stress function and data depth distortion using different sampling strategies, the lower the better (best results are shown in bold).

Dataset	RS	UDS	NUDS
Parkinson (stress)	3.42 ($\pm$2.55)	0.31 ($\pm$0.04)	**0.27** ($\pm$0.06)
Parkinson (depth distortion)	2.34 ($\pm$0.18)	1.88 ($\pm$0.05)	**1.81** ($\pm$0.23)
Stamps (stress)	3.58 ($\pm$0.69)	0.24 ($\pm$0.08)	**0.20** ($\pm$0.04)
Stamps (depth distortion)	5.24 ($\pm$0.39)	3.85 ($\pm$0.51)	**3.81** ($\pm$0.33)
Hepatitis (stress)	3.79 ($\pm$1.8)	0.22 ($\pm$0.06)	**0.14** ($\pm$0.05)
Hepatitis (depth distortion)	2.62 ($\pm$0.30)	1.94 ($\pm$0.31)	**1.87** ($\pm$0.13)

5.2. Uniform depth sampling (UDS)

In order to take into account the depth, we sample the computed depth values either uniformly or non-uniformly in the $[0, 1]$ depth range. Non-uniform sampling happens according to the depth distribution in the input space.

The first approach selects the control point in the original space by uniformly sampling its data depth. Afterwards the control points are positioned in the visual space again inside a disk with a diameter d_{max}. A control point with low depth value in the input

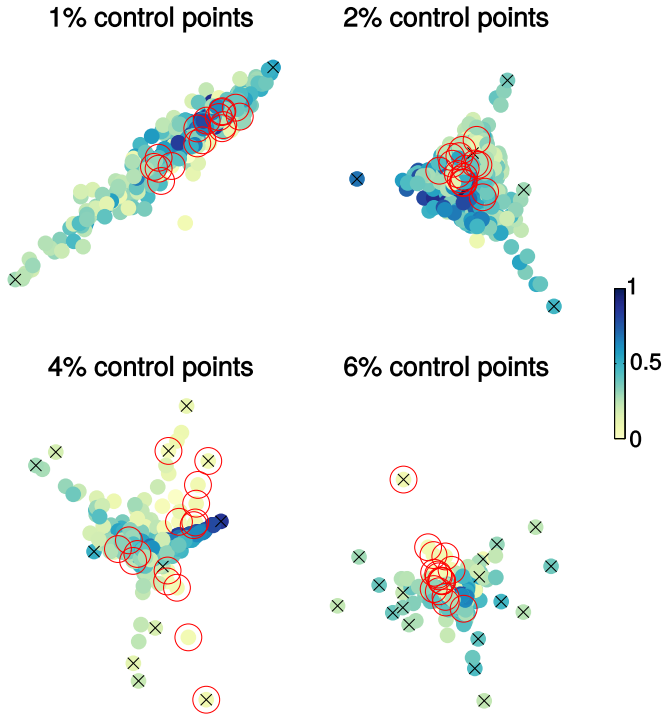

Fig. 10. Random sampling strategy applied to the Stamps dataset. Black marks depict the control points and red circles the outliers (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

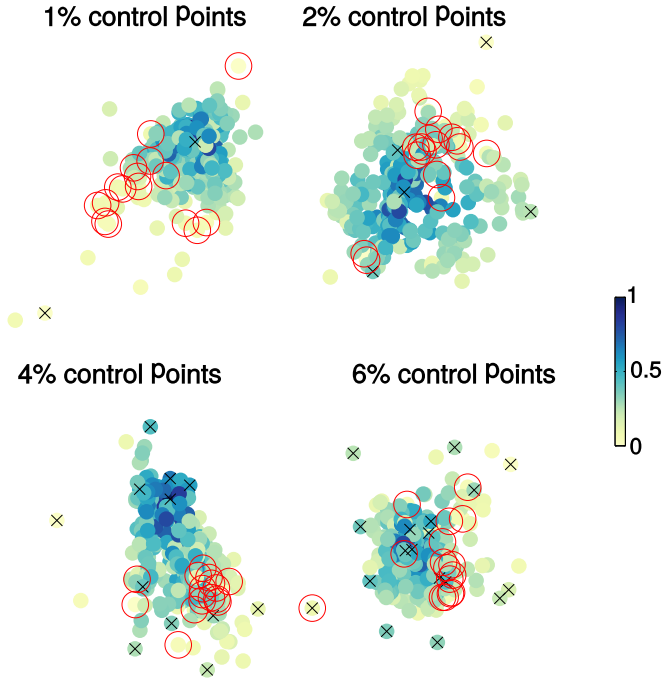

Fig. 11. Uniform depth sampling of the Stamps dataset.

space is positioned close to the disk center, whereas a control point with high depth value in the input space is positioned close do the disk perimeter.

Our formulation uses polar coordinates (ρ, θ), since with these coordinates our placement strategy is easy to specify. For the general position of the control points, the θ parameter is generated randomly following a uniform distribution in the $[0, 2\pi]$ range. We sort the control points according to their depth values in

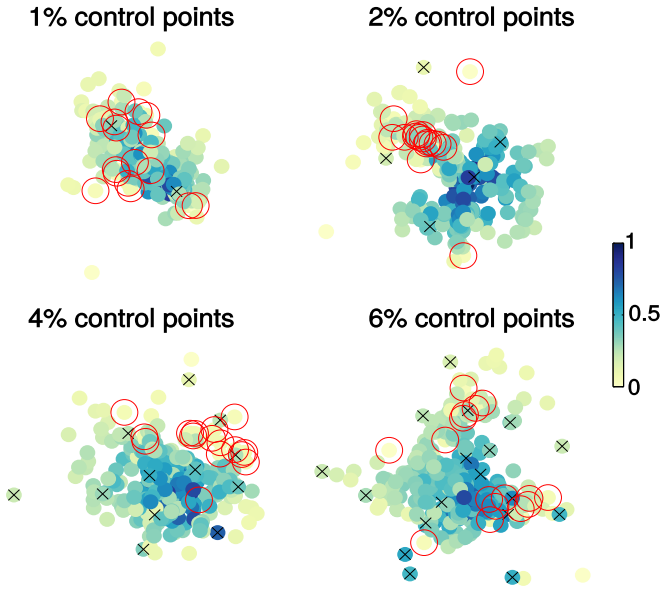

Fig. 12. Non-uniform depth sampling of the Stamps dataset.

descending order and map them to the radius ρ, uniformly sampled from $\left[0, \frac{d_{max}}{2}\right]$. The number of samples equals the number of control points. In Fig. 11 this strategy is illustrated with the Stamps dataset. Even using only 13 control points (4% of points), the projection was able to preserve the notion of continuity of the depth measure.

As can be seen in Table 4, the UDS approach improves the stress preservation compared to *RS*. Also it reduces the depth distortion, as can be expected from its radial construction, in which more central points on the input space are positioned on the 2D-disk center.

5.3. Non-uniform depth sampling (NUDS)

Our second step towards assessing the influence of sampling strategy is to use the data-depth distribution information to adapt the sampling. We use the empirical distribution function of the D^m to randomly generate a depth value v_i. For point selection, we pick the point with its data depth closest to v_i in the original space and use it as a control point for the projection. This process is done for the desired number of control points. For positioning the points on the visual space, their θ coordinates are generated identically to UDS case and their ρ coordinates are defined as their depth value in the original space.

The intuition of such weighting scheme is to mimic the depth distribution after the projection. In Fig. 12 this approach is compared against the uniform one.

As it can be seen in Table 4, it also improves upon RS both on stress and also on depth preservation, as expected. However, most of the results do not change considerably. While comparing results shown in Figs. 11 and 12, a slightly more regular distribution for intermediate depth values can be seen on the latter, for all tested number of control points.

5.4. Task-specific control point positioning

The UDS and NUDS sampling strategies use depth information to guide the projection process. While trying to preserve depth, general sampling strategies allow for other task-specific control-point layouts. We have performed an experiment, illustrated in Fig. 13, in which points with lowest depth values in the input space are placed as control points on the left side of the visual space,

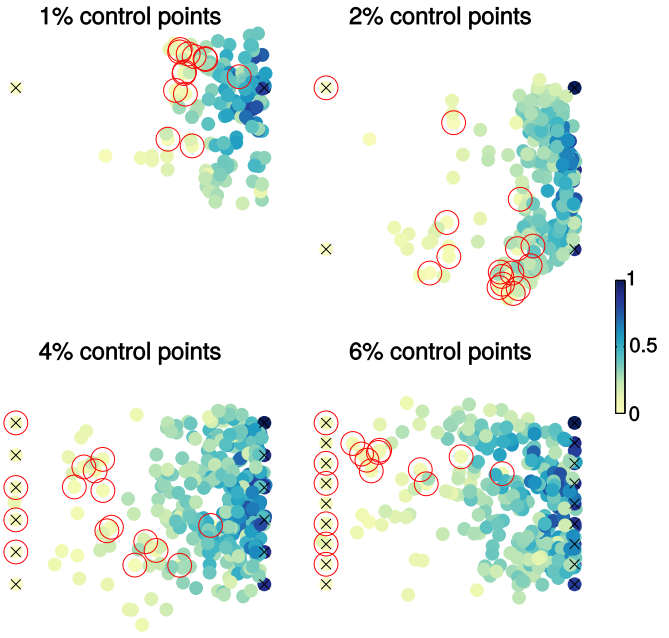

Fig. 13. Sampling of Stamps dataset with extrema depth values as control points.

while points with the highest depth values are placed on the right side.

The control points with the lowest depth values coincide with outliers, moving basically all the outliers towards the left side, even with as few as 12 control points (~4%). The main idea, of enforcing a placement of the data according to extrema depth values, is to explore the possibility of defining regions that are less likely of being contaminated with outliers. The experiment shows how increasing the number of control points impacts the outcome. The average stress achieved for this experiment is 0.22 ($\pm$0.041) and the average data-depth distortion is 4.11 ($\pm$0.73).

6. Discussion and limitations

For all datasets, Sammon mapping best preserved the data depth. The reason for this behavior is not clear yet.

While analyzing the different sampling strategies, two topics can be discussed: (a) the depth field opens possibilities for user interaction through different control-point layouts. The user could employ interactive tools, for specifying in which regions points with the extrema depth values should be placed, for instance. (b) Although the stress and the depth distortion originate from different motivations, in all performed experiments using control points they both showed a correlated behavior, i.e., both averages decreased simultaneously, on different scales though. While the stress value has been reduced by about 90%, the depth distortion has been reduced by about 30%, as compared to the random sampling strategy.

Fig. 6 indicates that using L_1D might produce poor results for complex datasets. In order to improve this result, we intend to explore more closely the Random Tukey depth [61], which uses subspaces to compute the depth in the input space. Initial results in Fig. 14 show that the Random Tukey depth is outperforming L_1D, as outliers have low depth values. This might be an interesting direction of further research.

Regarding the sampling strategies, there are not many differences between UDS and NUDS. Using the UDS strategy might be a better choice, since it is simpler. However, a scheme, non-linearly emphasizing depths close to the extremal values, might be interesting to explore. This could enforce a depth-based separation in the plane, similar to the experiment discussed in Section 5.4, but with even a better preservation of the depth distribution.

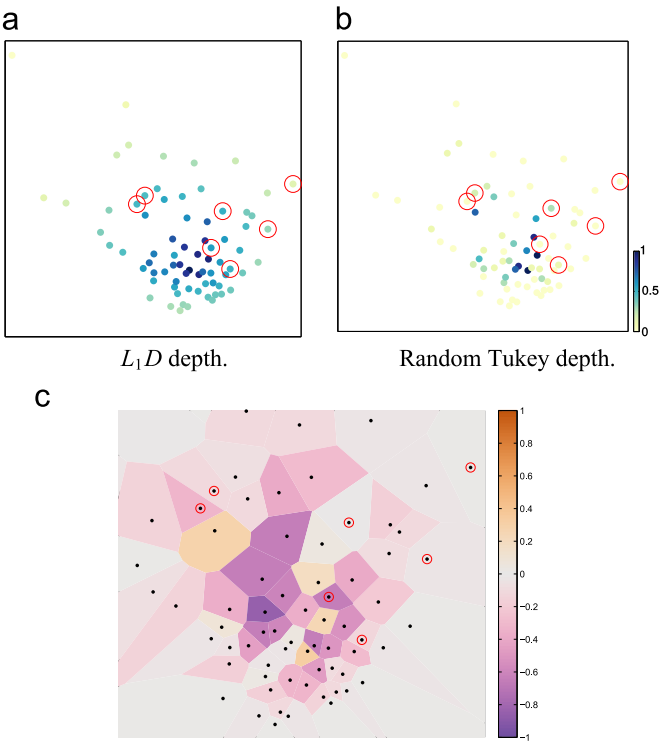

Fig. 14. Data depth computed for the Hepatitis dataset projected using Sammon mapping.

To the best of our knowledge, there is nothing specific in the literature to measure deviations of data depth. This has motivated the analysis with other measures like the stress and a neighborhood-based approach. Although they belong to a different class of distortion measures, their usage has been motivated by the possibility to explore relations with data depth preservation.

One interesting aspect of evaluating these three different strategies is that although LAMP does not explicitly intend to preserve on the projection any measure other than the distances locally, data depth preservation can be achieved by exploring control point layouts, even with few control points. The proposed strategies are ad hoc experiments in this direction, which can be improved by applying more refined optimizations.

One limitation of using data depth is that it does not support the analysis of clusters separately among the data, as well as class information, if available beforehand. However, the behavior of kmGMHD using Gaussian kernel in interesting, as a moderate depth value was distributed among different clusters, as seen in Fig. 2, which could lead to cluster analysis by properly interacting with the Hilbert space associated with the kernel.

Additionally, for multidimensional projection techniques based on kernel methods [62,63] the only available depth estimation tailored for such a scenario is the kmGMHD.

7. Conclusion and future works

In this paper we have proposed a novel approach for using order statistics as a quality measure for spatialization of multi-dimensional data. The computation of the data depth in the input space and in the visual space, after the data has been projected by

a MP technique, allow for evaluating centrality and depth preservation.

As pointed out by Tatu [13], there are few works investigating the relation of user perception and quality metrics. Within this context, user studies could be carried out in order to assess how centrality through data depth measures relate to human perception on scatter plots and multidimensional projections [9,24,34,35].

The global behavior of the data-depth distribution can be analyzed by using persistent homology with the framework proposed by Rieck and Leitte [16]. By doing so, one is able to explore in-depth differences between different depth functions (e.g., L_1D and Random Tukey depth). Additionally, the relation between depth functions and the co-ranking approach by Lee and Verleysen [29] is an interesting direction of investigation.

Recently, Pezzotti et al. [64] have proposed a hierarchical variation of SNE outlining an interesting direction of investigation on how it preserves the notion of data depth.

One important aspect, which can be further analyzed, is how central regions relate dimension-wise. In this work we did not address this problem. However, using linked-views of scatter plots and parallel coordinates might be a good starting point for such an analysis and would also allow for a better user interaction in the process.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable comments. The authors are supported by CNPq (fellowships #302643/2013-3 and #305796/2013-5) and São Paulo Research Foundation (FAPESP) grants #11/12263-0, #11/22749-8, #14/11296-0 and #14/09546-9.

References

[1] Joia P, Paulovich FV, Coimbra D, Cuminato JA, Nonato LG. Local affine multidimensional projection. IEEE Trans Vis Comput Graph 2011;17(12):2563–71 ISSN 10772626.
[2] Van Der Maaten L, Hinton G. Visualizing non-metric similarities in multiple maps. Mach Learn 2012;87(1):33–55 ISSN 08856125.
[3] Berger W, Piringer H, Filzmoser P, Groeller ME. Uncertainty-aware exploration of continuous parameter spaces using multivariate prediction. Comput Graph Forum 2011;30(3):911–20.
[4] Seo J, Shneiderman B. A rank-by-feature framework for interactive exploration of multidimensional data. Inf Vis 2005;4(2):96–113 ISSN 1473-8716.
[5] Turkay C. Integrating computational tools in interactive and visual methods for enhancing high-dimensional data and cluster analysis [Ph.D. thesis]. University of Bergen; 2013.
[6] Rauber PE, Feringa S, Celebi ME, Telea AC. Interactive image feature selection aided by dimensionality reduction. In: Proceedings of EuroVis workshop on visual analytics, 2015. p. 2–6.
[7] Martins RM, Coimbra DB, Minghim R, Telea AC. Visual analysis of dimensionality reduction quality for parameterized projections. Comput Graph 2014;41(1):26–42 ISSN 00978493.
[8] Bertini E, Tatu A, Keim D. Quality metrics in high-dimensional data visualization: an overview and systematization. IEEE Trans Vis Comput Graph 2011;17(12):2203–12 ISSN 10772626.
[9] Sedlmair M, Aupetit M. Data-driven evaluation of visual quality measures. Comput Graph Forum 2015;34(3):201–10 ISSN 01677055.
[10] Kulis B. Metric learning: a survey. Found Trends Mach Learn 2013;5:287–364 ISSN 1935-8237.
[11] Jeong DH, Ziemkiewicz C, Fisher B, Ribarsky W, Chang R. iPCA: an interactive system for PCA-based visual analytics. Comput Graph Forum 2009;28(3):767–74.
[12] Sedlmair M, Tatu A, Munzner T, Tory M. A taxonomy of visual cluster separation factors. Comput Graph Forum 2012;31(3):1335–44 ISSN 01677055.
[13] Tatu A. Visual analytics of patterns in high-dimensional data [Ph.D. thesis]. Universität Konstanz; 2013.
[14] Wilkinson L, Anand A, Grossman R. Graph-theoretic scagnostics, IEEE symposium on information visualization, INFOVIS, vol. 5, 2005. p. 157–64, ISSN 1522404X.
[15] Campos GO, Zimek A, Sander J, Campello RJGB, Micenková B, Schubert E, Assent I, Houle ME. On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study. Data Min Knowl Discov 2016;30(4):891–927.
[16] Rieck B, Leitte H. Persistent homology for the evaluation of dimensionality reduction schemes. Comput Graph Forum 2015;34(3):431–40 ISSN 01677055.
[17] Liu RY, Parelius JM, Singh K. Multivariate analysis by data depth: descriptive statistics, graphics and inference. Ann Stat 1999;27(3):783–858 ISSN 00905364.
[18] Serfling R. Depth functions in nonparametric multivariate inference. In: DIMACS series in discrete mathematics and theoretical computer science, vol. 72, 2006.
[19] Chandola V, Banerjee A, Kumar V. Anomaly detection: a survey. ACM Comput Surv 2009;41(3):15:1–58.
[20] Reimann C, Filzmoser P, Garrett R, Dutter R. Statistical data analysis explained: applied environmental statistics with R; 2011.
[21] Shapira L, Avidan S, Shamir A. Mode-detection via median-shift. In: IEEE 12th International Conference on Computer Vision, 2009. p. 1909–16.
[22] Fleishman S, Cohen-Or D, Silva CT. Robust moving least-squares fitting with sharp features. ACM Trans Graph 2005;24(3):544 ISSN 07300301.
[23] Chen C. Top 10 unsolved information visualization problems. IEEE Comput Graph Appl 2005;25(4):12–6 ISSN 02721716.
[24] Albuquerque G, Eisemann M, Magnor M. Perception-based visual quality measures. In: IEEE conference on visual analytics science and technology (VAST), 2011. p. 13–20.
[25] Aupetit M. Visualizing distortions and recovering topology in continuous projection techniques. Neurocomputing 2007;70(7–9):1304–30.
[26] Etemadpour R, Murray P, Forbes AG. Evaluating density-based motion for Big Data visual analytics. In: IEEE international conference on big data; 2014. p. 451–60.
[27] Schreck T, von Landesberger T, Bremm S. Techniques for precision-based visual analysis of projected data. Inf Vis 2010;9(3):181–93 ISSN 1473-8716.
[28] Lespinats S, Aupetit M. CheckViz: Sanity check and topological clues for linear and non-linear mappings. Comput Graph Forum 2011;30(1):113–25 ISSN 01677055.
[29] Lee JA, Verleysen M. Quality assessment of nonlinear dimensionality reduction based on K-ary neighborhoods. J Mach Learn Res: Workshop Conf Proc 2008;4:21–35.
[30] Venna J, Peltonen J, Nybo K, Aidos H, Kaski S. Information retrieval perspective to nonlinear dimensionality reduction for data visualization. J Mach Learn Res 2010;11:451–90.
[31] Heulot N, Aupetit M, Fekete J. Proxilens: interactive exploration of high-dimensional data using projections. In: Proceedings of the EuroVis workshop on visual analytics using multidimensional projections, 2013. p. 11–15.
[32] Martins RM, Minghim R, Telea A. Explaining neighborhood preservation for multidimensional projections. In: Computer graphics & visual computing (CGVC), 2015.
[33] Maier M, Luxburg UV, Hein M. Influence of graph construction on graph-based clustering measures. In: Advances in neural information processing systems, 2008. p. 1025–32.
[34] Etemadpour R, Linsen L, Crick C, Forbes A. A user-centric taxonomy for multidimensional data projection tasks. In: Proceedings of the international conference on information visualization theory and applications (IVAPP), Berlin, 2015. p. 11–14.
[35] Etemadpour R, Motta R, Paiva JGdS, Minghim R, de Oliveira MCF, Linsen L. Perception-based evaluation of projection methods for multidimensional data visualization. IEEE Trans Vis Comput Graph 2015;21(1):81–94.
[36] Hugg J, Rafalin E, Seyboth K, Souvaine D. An experimental study of old and new depth measures. In: Proceedings of the eighth workshop on algorithm engineering and experiments (ALENEX), 2006. p. 51–64.
[37] Rousseeuw PJ, Ruts I, Tukey JW. The Bagplot: a bivariate boxplot. Am Stat 1999;53(4):382–7.
[38] Zuo Y, Serfling R. General notions of statistical depth function. Ann Stat 2000;28(2):461–82 ISSN 00905364.
[39] Potter K, Rosen P, Johnson CR. From quantification to visualization: a taxonomy of uncertainty visualization approaches. IFIP Adv Inf Commun Technol 2012;377:226–47.
[40] Whitaker RT, Mirzargar M, Kirby RM. Contour boxplots: a method for characterizing uncertainty in feature sets from simulation ensembles. IEEE Trans Vis Comput Graph 2013;19(12):2713–22 ISSN 10772626.
[41] Mirzargar M, Whitaker RT, Kirby RM. Curve Boxplot: generalization of boxplot for ensembles of curves. IEEE Trans Vis Comput Graph 2014;20(12):2654–63 ISSN 1077-2626.
[42] Potter K, Kniss J, Riesenfeld R, Johnson CR. Visualizing summary statistics and uncertainty. Comput Graph Forum 2010;29(3):823–32 ISSN 01677055.
[43] Ding Y, Dang X, Peng H, Wilkins D. Robust clustering in high dimensional data using statistical depths. BMC Bioinforma 2007;8(Suppl 7):S8 ISSN 14712105.
[44] Small CG. A survey of multidimensional medians. Int Stat Rev 1990;58(3):263–77.
[45] Izem R, Rafalin E, Souvaine DL. Describing multivariate distributions with nonlinear variation using data depth 1. Technical Report. Tufts University, Department of Computer Science; 2008.
[46] Hu Y, Wang Y, Wu Y, Li Q, Hou C. Generalized Mahalanobis depth in the reproducing kernel Hilbert space. Stat Pap 2011;52(3):511–22 ISSN 09325026.
[47] Hofmann T, Schölkopf B, Smola AJ. Kernel methods in machine learning. Ann Stat 2008;36(3):1171–220.
[48] Härdle WK, Müller M, Sperlich S, Werwatz A. Nonparametric and semiparametric models. Springer Science & Business Media; 2004.
[49] Hoffmann H. Kernel PCA for novelty detection. Pattern Recognit 2007;40(3):863–74 ISSN 00313203.

[50] Vardi Y, Zhang C-H. The multivariate L_1-median and associated data depth. Proc Natl Acad Sci USA 2000;97(4):1423–6 ISSN 0027-8424.
[51] Won J-H, Lim J, Kim S-J, Rajaratnam B. Condition number regularized covariance estimation. J R Stat Soc Ser B, Stat Methodol 2013;75(3):427–50 ISSN 1369-7412.
[52] Maaten LVD, Hinton G. Visualizing Data using t-SNE. J Mach Learn Res 2008;9:2579–605 ISSN 02545330.
[53] Aupetit M, Catz T. High-dimensional labeled data analysis with topology representing graphs. Neurocomputing 2005;63:139–69.
[54] Harrower M, Brewer Ca. ColorBrewer.org: an online tool for selecting colour schemes for maps. Map Read: Theor Mapp Pract Cartogr Represent 2011;40 (1):261–8 ISSN 0008-7041.
[55] Lewis J, van der Maaten L, de Sa V. A behavorial investigation of dimensionality reduction. In: Proceedings of the cognitive science society, 2012. p. 671–6.
[56] Sammon JW. A nonlinear mapping for data structure analysis. IEEE Trans Comput 1969;C-18(5):401–9 ISSN 00189340.
[57] Hyvarinen A. Fast and robust fixed-point algorithm for independent component analysis. IEEE Trans Neural Net 1999;10(3):626–34.
[58] Limpert E, Stahel WA, Abbt M. Log-normal distributions across the sciences: keys and clues. BioScience 2001;51(5):341 ISSN 0006-3568.
[59] Lichman M. UCI machine learning repository. URL ⟨http://archive.ics.uci.edu/ml⟩; 2013.
[60] Tenenbaum JB. A global geometric framework for nonlinear dimensionality reduction. Science 2000;290(5500):2319–23 ISSN 00368075.
[61] Cuesta-Albertos J, Nieto-Reyes A. The random Tukey depth. Comput Stat Data Anal 2008;52(11):4979–88.
[62] Oglic D, Paurat D, Gärtner T. Interactive knowledge-based kernel pca. In: European Conference on Machine Learning and Knowledge Discovery in Databases, 2014. p. 501–16.
[63] Barbosa A, Paulovich F, Paiva A, Goldenstein S, Petronetto F, Nonato L. Visualizing and interacting with kernelized data. IEEE Trans Vis Comput Graph 2016;22(3):1314–25.
[64] Pezzotti N, Höllt T, Lelieveldt B, Eisemann E, Vilanova A. Hierarchical stochastic neighbor embedding. Comput Graph Forum 2016;35(3):21–30.